Weekly Report

Lu Junhua

2015 年 8 月 2 日

This week, I spared most of my time on Gongan Projects. We discussed with Prof. He on Monday, and went to TGRAM company the following four days.

According to He's suggestions, we extract out data firstly, and would implement Logistic regression on the data. Certainly, we will discuss again on improving the regression model to be better applicable on our data. The regression we did before is too "generalized" and needs revising. We should take the sampling method and normalization into consideration.

During data processing, we found many problems, many of which were not discovered by the company. Much of the details are recorded in our daily reports.

The following are brief summary of our work this week:

- Identify the data format. For tables will be extracted, they are static attributes table, hotel records table, internet cafe records table, crime records table. The last 3 table are extracted in this way:

  - 旅馆: 时长以天计算.

	2012.1	2012.2	…	…	…	2015.7
身份证 1	当月时长	当月时长				当月时长
身份证 2	当月时长	当月时长				当月时长
…						
…						
身份证 n	当月时长	当月时长				当月时长

  - 网吧: 类似上面的表, 但是以小时为单位计算.

  - 犯罪情况:

身份证	犯罪时间	犯罪类型
身份证号 1		
…		
身份证号 k		

- Ways to compute internet cafe and hotel records. It's not so easy since there are so many errors and noise in the data. We still need to improve the algorithm.

- Data sampling. Ke and Feng had many discussions with He, and temporarily we adopt an easy-understanding method to sample

- Re-import of data. We may make use all of the data later on so I learnt to use JDBC to make csv files into oracle database.

On the last day of business day, we extract 27390 permanent resident data, cost us 15 mins. It's fast, although there still exsits several errors in data, we may fix it as soon as possible. And once the data structure is determined further, I will write java codes which are corresponds to the data and import all of them into oracle database.

Besides, I had learned the method of k-n-Match problem. This method is easy to understand, and the algorithms are explicit. However, it can be used widely, not only in Gongan, netease project is also applicable. Since they are all Heterogeneous Data. However, whether adding timestamp as attributes is good or not remains discussion, I'm not sure about that, we may do more research on related fields.

I had read some materials related to the coding practice. I didn't focus on the details of implementation, but most on the algorithms behind them, like random greedy algorithm and squarified for treemap.